Digital Healthcare Empowering Europeans
R. Cornet et al. (Eds.)

doi:10.3233/978-1-61499-512-8-860

An original imputation technique of missing data for assessing exposure of newborns to perchlorate in drinking water

Alexandre CARON[a,c1], Guillaume CLEMENT[a,c], Christophe HEYMAN[a], Eva AERNOUT[a,c] , Emmanuel CHAZARD[c], Alain LE TERTRE[b]
[a] *Interregional epidemiology unit, Institute for public health surveillance; Lille, France*
[b] *French Institute for public health surveillance; Paris, France*
[c] *Department of Public Health; EA 2694, University of Lille; F-59000 Lille, France*

Abstract. Introduction. Incompleteness of epidemiological databases is a major drawback when it comes to analyzing data. We conceived an epidemiological study to assess the association between newborn thyroid function and the exposure to perchlorates found in the tap water of the mother's home. Only 9% of newborn's exposure to perchlorate was known. The aim of our study was to design, test and evaluate an original method for imputing perchlorate exposure of newborns based on their maternity of birth. **Methods**. In a first database, an exhaustive collection of newborn's thyroid function measured during a systematic neonatal screening was collected. In this database the municipality of residence of the newborn's mother was only available for 2012. Between 2004 and 2011, the closest data available was the municipality of the maternity of birth. Exposure was assessed using a second database which contained the perchlorate levels for each municipality. We computed the catchment area of every maternity ward based on the French nationwide exhaustive database of inpatient stay. Municipality, and consequently perchlorate exposure, was imputed by a weighted draw in the catchment area. Missing values for remaining covariates were imputed by chained equation. A linear mixture model was computed on each imputed dataset. We compared odds ratios (ORs) and 95% confidence intervals (95% CI) estimated on real versus imputed 2012 data. The same model was then carried out for the whole imputed database. **Results**. The ORs estimated on 36,695 observations by our multiple imputation method are comparable to the real 2012 data. On the 394,979 observations of the whole database, the ORs remain stable but the 95% CI tighten considerably. **Discussion**. The model estimates computed on imputed data are similar to those calculated on real data. The main advantage of multiple imputation is to provide unbiased estimate of the ORs while maintaining their variances. Thus, our method will be used to increase the statistical power of future studies by including all 394,979 newborns.

Keywords. Perchlorates, Epidemiologic Studies, Imputation, Missing data.

Introduction

Perchlorate salts are inorganic compounds containing the perchlorate anion. They are mainly used as propellant in rocket fuels, explosives or in airbag trigger system. When

[1] Corresponding Author : alexandre.caron-2@univ-lille2.fr

released in groundwater, perchlorate anions stay unreactive for a long period of time. Humans can be exposed by eating food and drinking water that contains perchlorate. At doses incommensurate with environmental exposure, perchlorate inhibits the uptake of iodine by the thyroid by competing with iodine absorption in the Natrium-Iodine Symporter (NIS) leading to relative hypothyroidism. At environmental doses, some studies suggest the possibility of biological disorders, especially in women with iodine deficiency [1,2]. Pregnant women are more sensitive to altered uptake of iodine due to an increased clearance and fetal needs. In the case of decreases in iodine uptake, the thyroid can no longer compensate [3,4]. A similar mechanism may be observed on the fetal thyroid since perchlorate actively passes through the placental barrier [5,6]. Thus, small changes in maternal or fetal thyroid function could cause long-term effects on the neurological development of the child [7], such as autism, attention disorders, hyperactivity or decreased intelligence quotient (IQ) [8–10].

Perchlorate has been recently detected in tap water samples of the Nord Pas-de-Calais (NPDC), a region located in the North of France. Following this discovery, we set up an epidemiological study to assess the association between the concentration of perchlorate in the tap water supply of the mother's home and newborn thyroid function. Thyroid function of all 394,979 newborns included in the study was evaluated by thyroid stimulating hormone (TSH) level after 3 days of life. Newborns were exposed to perchlorate during pregnancy through tap water. Perchlorate exposure was linked to newborn's TSH level using municipality of residence of the newborn's mother. Unfortunately, exposure was unknown because of a lack of data entry. The municipality of residence of the newborn's mother was not systematically computerized before 2012. Thus, only 9% of the database could have been analyzed, leading to a major loss of statistical power.

Although many statistical methods exist to deal with incomplete datasets, their use is not yet systematic. Multiple imputation is the best method currently available [11]. Nowadays, computation time is no longer prohibitive and methods are incorporated into most statistical software. Therefore these techniques can be used in the epidemiological field. Multiple Imputation by Chained Equation is well-known method and can easily be performed with the statistical software R and its MICE package [12]. Its imputation algorithm uses the other variables to predict the missing value through multiple regressions. It can only be used under the missing completely at random assumption (MCAR). Moreover, this technique requires that less than 15% of data are missing. The municipality of residence of the newborn's mothers was MCAR (not computerized). However the use of this "classical" method was here limited by the high percentage of missing geographic data before 2012 and the large number of possible locations. Indeed, the algorithm had to predict the birthplace from 1,545 municipalities. Furthermore, no other variable in the first dataset appeared to predict the municipality of residence of the newborn's mother.

The aim of our study was to design, test and evaluate an original method for imputing the municipality of residence of the newborn's mothers (and as a result the perchlorate exposure) based on their maternity of birth.

1. Methods

1.1. Material

The study population consisted of all newborns in NPDC maternity wards between December 1, 2004 and October 16, 2012, and whose mother was residing in one of the municipalities of the region aforementioned (N=394,979). This first database was an exhaustive collection of newborn's thyroid function measured during a systematic neonatal screening in NPDC. We reused TSH level and other individual variables for our study. The second database contained tap water supply concentration of perchlorate, carried out in each of the 1,545 municipalities of NPDC. For each birth, the following variables were considered: TSH level, individual adjustment variables (term, birth weight, gender, date of birth, age at the time of sampling), and the maternity of birth. Exposure was defined as the concentration of perchlorate in the tap water supply of the municipality of residence of the mother's home during her pregnancy. This data was only available for the year 2012 (N=36,695) which represent 9.28% of database.

The PMSI database is a French nationwide exhaustive database of inpatient stays derived from the medical insurance payment system. For each inpatient stay, this database describes notably the Diagnosis Related Group (DRG). We reused it to extract all births in NPDC maternity over the study period. DRG of "vaginal" and "cesarean" deliveries enabled us to find all birth over the study period. Thus, we were able to determine the catchment area of the maternity wards. We computed the probability of birth in a municipality X for a given maternity ward as the number of births from municipality X divided by the total of birth in the maternity ward.

1.2. Missing Data Imputation Method

Imputation and statistical analysis were performed with R version 3.1.1 [13]. From the maternity ward of birth, we randomly selected a municipality of newborn's mother weighted by the probabilities calculated from the catchment area. Thus, we were able to associate a concentration of perchlorate for each birth in the database. In the same way, we performed 20 draws from the database of perchlorate levels. Hence we obtain 20 concentrations of perchlorate for each birth. In the meantime, we imputed the missing data for gender, birth weight and term via chained equations method using MICE package [12]. The maximum number of iterations for the Gibbs sampler was limited to 20. We then merged each perchlorate concentration draw with the corresponding imputed dataset inside a *multiply imputed dataset* (class .mids of MICE package).

We carried out the same model on the 20 datasets: a multivariate linear regression model studying the relationship between perchlorate exposure and newborn TSH. We adjusted on individual variables, and add a random effect on municipality of residence (a unique exposure level by municipality lead to serial correlation of residuals). The resulting models only differed by their imputed values. The regression coefficients were pooled and the variability produced by the imputation was automatically taken into account by Rubin's rules.

We evaluated the method by comparing the coefficients and 95% confidence intervals (95% CI) calculated on the 2012 real data (Gold Standard Model) versus 2012 with imputation of the municipality of residence (Imputed Model 1). Parameters were also computed for the whole 2004-2012 imputed database (Imputed Model 2).

2. Results

2.1. Descriptive Analysis of Imputed Data

We included 36,695 newborns in 2012 and 357,772 newborns between 2004 and 2011. Mean of gold standard exposure is 4.87 µg/l (median = 3.00). Perchlorate levels reach a maximum 77.00 µg/l. Multiple imputation of perchlorate exposure was performed creating twenty different distributions of perchlorate exposure. A graphically similar distribution was observed between the Gold Standard and Imputed Model 1 (Figure 1). Means range from 4.83 µg/l to 4.94 µg/l. Quartiles are identical in all imputed and real datasets. Three imputed dataset had a maximum of 67.00 µg/l.

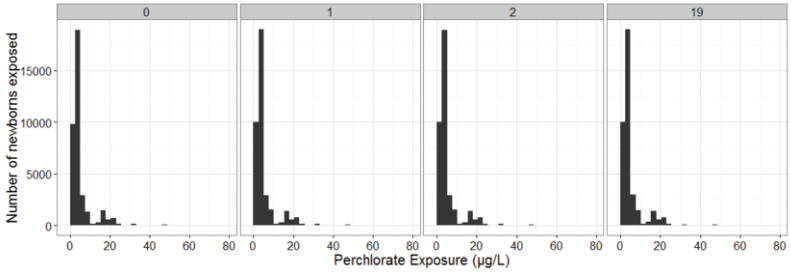

Figure 1. Set of 3 randomly selected exposure imputations (1, 2, 19) and comparison to Gold Standard (0)

2.2. Comparison of Multivariate Analysis Coefficients and Confidence Intervals

In order to assess the accuracy of the imputation model, we compared odds ratios between Gold Standard and Imputed Model 1 (36,695 newborns). The ORs imputed from Model 1 show the same trends as the ORs of the Gold Standard (Table I). Variables number 4 and 5 are not statistically significant. The amplitude of the 95% CI is preserved between both models, which was confirmed by similar test statistics (t) for each model. When increasing the statistical power as in Imputed Model 2 (357,772 newborns), the ORs remain stable and the 95% CI tighten considerably (Table 1).

Table 1. Odds ratios and 95% confidence intervals of Gold Standard and Imputed Models.

	Gold Standard (2012)		Imputed Model 1 (2012)		Imputed Model 2 (2004-2012)	
1	**1.047**	[1.029 ; 1.065]	**1.042**	[1.024 ; 1.059]	**1.041**	[1.035 ; 1.047]
2	**1.086**	[1.0790 ; 1.094]	**1.084**	[1.076 ; 1.090]	**1.077**	[1.074 ; 1.079]
3	**0.99992**	[0.99990 ; 0.99994]	**0.99993**	[0.99991 ; 0.99995]	**0.99993**	[0.99993 ; 0.99994]
4	**1.0007**	[0.9989 ; 1.0026]	**0.9990**	[0.9971 ; 1.0008]	**1.0017**	[1.0012 ; 1.0022]
5	**0.999**	[0.985 ; 1.014]	**0.998**	[0.986 ; 1.007]	**1.002**	[0.998 ; 1.006]
6	**0.979**	[0.961 ; 0.997]	**0.979**	[0.961 ; 0.997]	**0.947**	[0.941 ; 0.953]

3. Discussion

In this study, we used an original imputation method to impute data in an epidemiological database. The model estimates computed on imputed data were similar to those calculated on real data, but with considerably tightened 95% CI.

Multiple imputation provides unbiased estimates of the OR while maintaining its variance (and the confidence interval) under the missing completely at random (MCAR) assumption. The probability of an observation being MCAR is independent of any observable or unobservable variable. The latter assumption is verified in our database as municipality was not computerized before 2012. Chained equation could not be used as is in our study. Indeed, this method needs the other variable in the dataset to be predictor of the imputed variable. Using term, birth weight, gender or date of birth as predictor of maternity of birth would have been inappropriate. Catchment area was therefore an interesting alternative to impute the municipality of the mother. The slight loss of power related to imputation procedure is largely compensated by the increased sample size by ten.

Although 36,695 is already a large epidemiological sample. However, the prevalence of thyroid dysfunction in newborn is less than 1/3,000. Therefore, a high number of births is required to get sufficient statistical power. Running multiple imputation on 20 datasets with more than 300,000 rows could have been computationally prohibitive some time ago. Nevertheless, computing power is continually improving and it was not an issue for our study. This method will be used to increase the statistical power by including newborns with imputed data for the municipality of the mother and consolidate our epidemiological conclusions.

References

[1] B. C. Blount, J. L. Pirkle, J. D. Osterloh, L. Valentin-Blasini, and K. L. Caldwell, "Urinary perchlorate and thyroid hormone levels in adolescent and adult men and women living in the United States," *Environ. Health Perspect.*, vol. 114, no. 12, pp. 1865–1871, Dec. 2006.

[2] M. B. Zimmermann, "Iodine deficiency," *Endocr. Rev.*, vol. 30, no. 4, pp. 376–408, Jun. 2009.

[3] C. R. Fantz, S. Dagogo-Jack, J. H. Ladenson, and A. M. Gronowski, "Thyroid function during pregnancy," *Clin. Chem.*, vol. 45, no. 12, pp. 2250–2258, Dec. 1999.

[4] D. Glinoer, "The regulation of thyroid function during normal pregnancy: importance of the iodine nutrition status," *Best Pract. Res. Clin. Endocrinol. Metab.*, vol. 18, no. 2, pp. 133–152, Jun. 2004.

[5] J. Logothetopoulos and R. F. Scott, "Active iodide transport across the placenta of the guinea-pig, rabbit and rat," *J. Physiol.*, vol. 132, no. 2, pp. 365–371, May 1956.

[6] O. Dohán, C. Portulano, C. Basquin, A. Reyna-Neyra, L. M. Amzel, and N. Carrasco, "The Na+/I symporter (NIS) mediates electroneutral active transport of the environmental pollutant perchlorate," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 51, pp. 20250–20255, Dec. 2007.

[7] G. R. Williams, "Neurodevelopmental and neurophysiological actions of thyroid hormone," *J. Neuroendocrinol.*, vol. 20, no. 6, pp. 784–794, Jun. 2008.

[8] G. C. Román, "Autism: transient in utero hypothyroxinemia related to maternal flavonoid ingestion during pregnancy and to other environmental antithyroid agents," *J. Neurol. Sci.*, vol. 262, no. 1–2, pp. 15–26, Nov. 2007.

[9] J. E. Haddow, G. E. Palomaki, W. C. Allan, J. R. Williams, G. J. Knight, J. Gagnon, C. E. O'Heir, M. L. Mitchell, R. J. Hermos, S. E. Waisbren, J. D. Faix, and R. Z. Klein, "Maternal thyroid deficiency during pregnancy and subsequent neuropsychological development of the child," *N. Engl. J. Med.*, vol. 341, no. 8, pp. 549–555, Aug. 1999.

[10] F. Vermiglio, V. P. Lo Presti, M. Moleti, M. Sidoti, G. Tortorella, G. Scaffidi, M. G. Castagna, F. Mattina, M. A. Violi, A. Crisà, A. Artemisia, and F. Trimarchi, "Attention deficit and hyperactivity disorders in the offspring of mothers exposed to mild-moderate iodine deficiency: a possible novel iodine deficiency disorder in developed countries," *J. Clin. Endocrinol. Metab.*, vol. 89, no. 12, pp. 6054–6060, Dec. 2004.

[11] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 2009.

[12] S. Buuren and K. Groothuis-Oudshoorn, "MICE: Multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, no. 3, 2011.

[13] *R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.*